Supplementary Methods

Re-analysis of copy number data

Raw copy number files were downloaded from the TCGA data portal (http://cancergenome.nih.gov/dataportal). The data were submitted by the following four research groups using three assay platforms: 1) Broad Institute of MIT and Harvard (Broad) using Affymetrix Genome-wide Human SNP Array 6.0; 2) Memorial Sloan-Kettering Cancer Center (MSKCC) using Agilent Human Genome CGH MicroArray 244A; 3) Harvard Medical School using Agilent Human Genome CGH MicroArray 244A; and 4) Stanford University School of Medicine using Illumina 550K Infinium HumanHap500 SNP Chip.

For the Affymetrix SNP 6.0 data, the copy number data were re-computed from the *.CEL files using Affymetrix Genotyping Console software (version 2.1). The default parameters were used and HapMap270.425 model was the reference.

For the other three data sets, the copy number data of the original submission were used.

For all four sets of copy number data, we ran DNAcopy (v 1.12.0), an R package that implements the Circular Binary Segmentation (CBS) algorithm [1] to determine the segments of deletion/amplification. The default parameters were used for all data sets. The input was the log2 ratio of each marker; where there was matching tumor/normal copy number data, the log2 ratio of the tumor/normal pair was used, otherwise the sample/reference log2 ratio was used.

The segments of deletion/amplification generated by CBS were then used to build the copy number heat maps. Each segment was mapped initially to a high-resolution set of fixed-size (10,000 base pairs) genomic bins. For each bin, the total number of samples showing amplification or loss in the bin was tallied. The high-resolution bins were then mapped to a set of low-resolution (200,000 base pairs) bins. The copy number value of each low-resolution bin was taken from the contained high-resolution bin that showed the greatest number of amplifications or losses across samples. For the heat map display, the continuous copy number values computed by the CBS algorithm were scaled to 4 discrete levels of loss (< 0.3, < 0.8, < 1.3, < 1.8), a neutral level (> 1.8 and <= 2.2), and 5 discrete levels of gain (<= 2.7, <= 3.2, <= 3.7, <= 4.2, > 4.21). Tumors with paired normal were separated from those with no matching normal. Amplifications and deletions found in normal samples were considered copy number polymorphisms if the frequency is >= 10%.

A similar protocol was used to generate the input data for the "landscape" view of the Cancer Genome Workbench (CGWB). The low-resolution bin size was set to 1Mb to match the scale of chromosome ideogram on the display.

The genome view of the copy number data on CGWB (displayed as part of TCGA_GBM_CopyNumber and TCGA_GBM_Integrated tracks) shows CBS output with no post-processing. Only one copy number data source is selected for each sample. For samples that have copy number data generated by multiple groups, copy number data are selected in the following order of precedence: Broad, MSKCC, Harvard and Stanford.

Exon array data

Exon array expression data were downloaded from the TCGA data portal. The data were generated using the Affymetrix Human Exon 1.0 ST Array by Lawrence Berkley National Laboratory. Exons that do not overlap with RefSeq (http://www.ncbi.nih.gov) were filtered out. Genomic locations of RefSeq exons were obtained from the RefFlat table of the UCSC genome browser (http://genome.ucsc.edu). The under- and over-expression display is calculated based on the deviation from the mean expression value of each exon.

Putative somatic mutation analysis

TCGA traces were downloaded from the NCBI's trace archive (http://www.ncbi.nlm.nih.gov). Variations were analyzed using the SNPdetector [2] and IndelDetector pipeline[3]. To distinguish genetic variations from paralogous variations due to non-specific PCR primers, a module for identifying non-specific primers was implemented as follows. For any amplicon that shares >90% identity with a second genomic locus, sequence variations are compared with paralogous variations. If the forward and the reverse primer sequences are identical at the paralog or if the majority of the sequence variations are identical to paralogous variations, the primer pair for the amplicon is considered non-specific and all variations are discarded.

Validated mutations

The chromosome locations and genotypes of validated somatic mutations were extracted from the MAF files on the TCGA data portal. The data were submitted by the following three genome sequencing centers: Broad Institute of MIT and Harvard, Washington University, and Baylor College of Medicine. For tumor samples with conflicting genotypes reported by different centers, the homozygous genotype is preferred where there is also somatic copy number variation; otherwise, the heterozygous genotype is selected.

Somatic mutation annotation

Putative variations as well as validated somatic mutations were mapped to RefSeq to compute the amino acid changes and location on the 3D structure as described previously[3]. The impact of the protein alteration on conserved domains is assessed by logE[4] and SIFT[5] methods.

Methylation data

Methylation data, submitted by Johns Hopkins/University of Southern California, were downloaded from the TCGA data portal. The data were generated using the Illumina DNA Methylation OMA002 and OMA003 Cancer Panels. Genomic location of a methylation site was computed by running a search of its probe sequence against the reference human genome using the program BLAT [6]. Methylation beta values are displayed on CGWB at 3 discrete levels (<0.25, <0.75, >=0.75) to represent under-, neutral, and over-methylation.

Gene expression analysis

TCGA Gene expression data, submitted by the Broad Institute, were downloaded from the TCGA data portal. The data were generated using Affymetrix HT Human Genome U133A chip. Only GBM tumor expression data are available in this data set. To obtain non-tumor reference expression data, gene expression data generated by Li et al [7] using Affymetrix U133 Plus2 were downloaded from the Repository of Molecular Brain Neoplasia Data web site ("Rembrandt", http://caintegrator-info.nci.nih.gov/rembrandt). The "Rembrandt" CEL files were converted into the U133A format and the two data sets were normalized using RMA (Robust MultiChip Average) with custom CDF (Chip Definition File) that removes the non-specific and mis-targeted probes to create gene-specific expression values[8]. The non-tumor data in Li et al provides the baseline expression value. The probability of each gene being in the 'up' state for each sample was then calculated by fitting the data to a mixture of 2 gamma distributions[9]. Clustering using only the top 1000 most variable genes separated normal from tumor samples with no apparent batch effect.

Correlation analysis

TCGA has microRNA (miRNA) and gene expression data from the same samples, which allows us to examine the possible regulation of gene expression by microRNA expression. A negative correlation between the expression of microRNA and its target genes (miRNA-mRNA correlation) is expected. To explore this, we download the gene expression and microRNA expression data (UNC Agilent G4502A_07 and UNC Agilent 8x15K) directly

from the TCGA Data Portal (http://tcga-data.nci.nih.gov/tcga/homepage.htm). The miRNA target site data was downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables?org=Human&db=hg18&hgsid=145859044&hgta_doMainPage=1&hgta_group=regulation&hgta_track=targetScanS&hgta_table=targetScanS), which is predicted by TargetScan[10]. The Pearson correlation analysis was performed for each miRNA and predicted gene pair: there are a total of 122,794 pairs, 99,864 pairs for TCGA GBM and TCGA OV, respectively). The p-value for the correlation was adjusted for multiple testing. A miRNA-mRNA pair with significant correlation was then defined by meeting the following two criteria: 1) correlation coefficient greater than 2 standard deviations from the mean; and 2) the adjusted p-value less than 0.001. Consequently, 2,903 and 884 significant negative correlations were found between miRNAs and their target genes in TCGA GBM and TCGA OV, respectively.

12295 genes in TCGA GBM and 12151 genes in TCGA OV have both expression and methylation data. Methylation beta value and log2 intensity of gene expression data described above were used to calculate correlation-coefficient using the Perl module Math-NumberCruncher-5.00.

Normalized gene expression and methylation data for TCGA glioblastoma multiforme (GBM) and ovarian cancer (OV) were obtained from the TCGA Data Portal (http://tcga-data.nci.nih.gov). Gene expression analyses of TCGA samples using Agilent Expression 244K microarrays were performed by The University of North Carolina at Chapel Hill. Methylation profilings were carried out by Johns Hopkins University and University of Southern California. For TCGA GBM samples, Illumina Golden Gate Methylation Bead array, customized Golden Gate Methylation Bead array, and Illumina Infinium Human DNA Methylation 27 platform were used. For TCGA OV samples, Illumina Infinium Human DNA Methylation 27 platform was used.

Next-generation sequencing data

BAM files, which record the alignment between each next-gen read and the reference human genome (hg18), for 19 exon capture and 5 whole-genome sequencing data of TCGA Ovarian were downloaded from dbGaP (http://www.ncbi.nlm.nih.gov/gap/). Read coverage for each genomic base was calculated by including only reads with quality score >=15.

These BAM files were also used to find substitution variations (i.e. SNPs). The SNP finder uses the "Picard" Java API (http://picard.sourceforge.net/index.shtml) to read SAM data[11].

SNPs are detected by iterating through one or more .bam files, which are typically pre-sorted by reference sequence name and mapping position. Multiple files (e.g. for a tumor/normal pair) may be analyzed together; in this case a wrapper is used which combines the iterators from all files, essentially creating a single virtual file which returns reads from multiple sources in one ordered stream.

Putative SNP sites are detected by comparing non-clipped read alignments to the reference sequences. Each SAM record's "CIGAR" (Concise Idiosyncratic Gapped Alignment Report) alignment field is also parsed to determine putative indel sites. These observations are aggregated and periodically evaluated once all reads covering a section of the genome have been read. Various heuristics are applied to each aggregation of putative sites to determine whether a final SNP or indel call should be made.

All reads pass through a mismatch filter, which is intended to prevent potentially mismapped reads from contributing to SNP or indel calling. Currently we exclude any read from consideration having more than 2 mismatches of quality 15 or better, or more than 4 mismatches of any quality. Mismatches are not included in this count if they occur in soft-clipped regions, or within poly-X regions of 5 nt or longer.

For SNPs, the alternative allele is the most frequently-observed non-reference allele (this may be the dominant or only observed allele for samples homozygous for the non-

reference allele). A minimum nucleotide quality of 15 is required at each variant site. For SNPs, this is the quality at the SNP base. Additionally a minimum amount of high-quality flanking sequence may be required (default is 5 or more nt of quality 15+). Simple filters are provided to control minimum required observation frequencies: minimum read coverage, minimum number of reads supporting the alternative allele, and minimum observed frequency of the alternative allele. Reads with SAM's "PCR/Optical duplicate" flag set are excluded from the analysis, as these typically represent monoclonal artifacts. Two additional filters help combat similar effects: one which requires at least one sequence at the variant site to have 10 nt or more of flanking sequence, and another which requires at least 2 unique mapping positions for reads which contribute to a variant. All the filters described above are configurable; different parameters may be appropriate depending on the desired analysis.

References

1. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-63 (2007).
2. Zhang, J. et al. SNPdetector: A Software Tool for Sensitive and Accurate SNP Detection. *PLoS Comput. Biol.* **1**, e53 (2005).
3. Zhang, J. et al. Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB). *Genome Res.* **17**, 1111-7 (2007).
4. Clifford, R.J., Edmonson, M.N., Nguyen, C. & Buetow, K.H. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* **20**, 1006-14 (2004).
5. Ng, P.C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11**, 863-74 (2001).
6. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-64 (2002).
7. Li, A. et al. Genomic changes and gene expression profiles reveal that established glioma cell lines are poorly representative of primary human gliomas. *Mol. Cancer Res.* **6**, 21-30 (2008).
8. Zhang, J., Finney, R.P., Clifford, R.J., Derr, L.K. & Buetow, K.H. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics* **85**, 297-308 (2005).
9. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).
10. Efroni, S., Carmel, L., Schaefer, C.F. & Buetow, K.H. Superposition of transcriptional behaviors determines gene state. *PLoS One* **3**, e2901 (2008).
11. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).